On Mixture Memory GARCH Models

Muyi Li

School of Economics & Wan Yanan Institute for Studies in Economics

Xiamen University, Xiamen, China

Wai Keung Li and Guodong Li

Department of Statistics and Actuarial Science

University of Hong Kong, Pokfulam Road, Hong Kong

Abstract

We propose a new volatility model, which is called the mixture memory GARCH (MM-GARCH) model. The MM-GARCH model has two mixture components, of which one is a short memory GARCH and the other is the long memory FIGARCH. The new model, a special ARCH(∞) process with random coefficients, possesses both the properties of long memory volatility and covariance stationarity. The existence of its stationary solution is discussed. A dynamic mixture of the proposed model is also introduced. Other issues, such as the EM algorithm as a parameter estimation procedure, the observed information matrix which is relevant in calculating the theoretical standard errors, and a model selection criterion are also investigated. Monte Carlo experiments demonstrate our theoretical findings. Empirical application of the MM-GARCH model to the daily S&P 500 index illustrates its capabilities.

Key words: long memory in volatility, covariance stationarity, mixture ARCH(∞), EM algorithm.

1 Introduction

Long range dependence (LRD) in the squares of speculative returns has been widely accepted as a stylized fact since it is firstly reported by Ding et al. (1993). Among numerous long memory volatility models, the FIGARCH model (Baillie et al., 1996) and the HYGARCH model (Davidson, 2004) are the most popular parametric specifications. In comparison with the ARCH(p) process of Engle (1982) and the GARCH(p,q) process of Bollerslev (1986) which have geometric decaying coefficients of the squared residuals, both of the FIGARCH and the HYGARCH have hyperbolic decaying counterparts. Different from the FIGARCH, the variance of which always does not exist, the HYGARCH model releases the unit-amplitude restriction hence making it possible to have both characters of long memory volatility and covariance stationarity. Besides the HYGARCH, there are some other long memory volatility models that possess these two desired properties, see Zaffaroni (2004) and Robinson and Zaffaroni (2006).

Before starting further investigation on the memory effect of speculative returns, a few remarks on the definition of long memory volatility are needed. First, under the framework of ARCH(∞) processes (Robinson, 1991), the decaying structure of the autocorrelation function (ACF) fully depends on the impulse response coefficients (IRFs). That is, if the coefficients have the hyperbolic decaying rates, so do the ACFs. Second, an important fact to be noted is that the squares of an ARCH(∞) process with a finite fourth moment always have short memory in the sense of absolutely summable autocovariances (Giraitis and Surgailis, 2002). In this paper, processes of long memory in volatility are referred to those whose ACFs of squares or IRFs decay hyperbolically. Compared to this, processes having the geometric decaying ACFs or IRFs belong to the family of short memory volatility processes.

Since first introduced by Davidson (2004), the HYGARCH model has attracted more attentions in its empirical applications. In the theoretical aspect, Conrad (2010) gives the non-negativity conditions of the conditional variance, and Li et al. (2011) developed the score test for geometric decay against hyperbolic decay. Note that the conditional variance of the HYGARCH process can be interpreted as the weighted summation of those of a common GARCH and a FIGARCH model respectively with weights α and $1-\alpha$, see Li et al. (2011). The idea of weighting two ARCH-type models can be dated back to Ding and Granger (1996), where a common GARCH model and an IGARCH model are employed to form a new ARCH(∞) model. Although the HYGARCH model is well motivated, it has met several difficulties in empirical studies. For example, when Davidson (2004) applied the HYGARCH to model the volatility dynamics of some Asian currencies during the Asian crisis in 1997-1998, many of these series were found to have their variance indicator $\alpha > 1$, which implies that the series is not stationary in terms of the second order moment. Another problem is the use of a single regime to cover a long period is not strongly convincing. There are plentiful of works aiming to discover the connection between non-stationarities and long run dependence. For the sake of brevity, we focus on a few examples. Ding and Granger (1996) claimed that different volatility processes should be used for different period. Mikosch and Starica (2004) show theoretically that the IGARCH effect could be due to the behavior of the estimator under model misspecification, and the combination of different short memory GARCH processes can result in long memory phenomenon in the sample ACF. Baillie and Morana (2009) pointed out that the existence of structural change in the conditional variance might be a plausible source of long memory volatility, hence they allowed the intercept in the conditional variance to be time varying in order to provide the model with more flexibility in the volatility structure. Based on the literature abovementioned and references therein, we consider a regime switching model by interpreting the weight α in the HYGARCH model as a regime indicator. This kind of regime switching model is also referred to as mixture models, where the switching probabilities are allowed to depend only on the lagged observed variables. Some seminal articles on mixture time series models are Wong and Li (2000, 2001a,b). Cheng et al. (2009) employed a two-regime GARCH process for the conditional variance and adopted a logistic specification for the mixture proportion. Distinguishing from the above models where the mixture components have the same common GARCH structure, the mixture model proposed here has the following structure. One of the components is a common GARCH (the short memory GARCH) and the other component is a FIGARCH (the long memory GARCH). Therefore, we call it the mixture memory GARCH (MM-GARCH) model. Similarly, the mixture probability can be constant or dynamic depending on economic theories or practical requirements. In the latter case, the model is called a dynamic MM-GARCH model. The mixture probability makes it possible for one of the components assuming a non-stationary ARCH(∞) process while the whole time series can still be second-order stationary. At the same time, it enjoys hyperbolically decaying IRFs just as the FIGARCH process since the hyperbolic lags will dominate the geometric ones when the lags are large enough.

The paper is organized as follows. Section 2 introduces the MM-GARCH model and studies the conditions of stationary solutions with the second and higher order moments. We also introduce a more general dynamic MM-GARCH model. In Section 3, we discuss the estimation procedure based on the EM algorithm and a simple version of the observed information matrix is provided. The model selection criterion is briefly discussed in Section 4. Section 5 demonstrates the finite-sample performance of the MM-GARCH models by Monte Carlo experiments. An example of value-at-risk (VaR) tests illustrates the capacity of the new models on forecasting the volatility dynamics in section 6. Proofs and technical details are given in the Appendix.

2 Mixture Memory GARCH Models

The mixture memory GARCH (MM-GARCH) process $\{e_t\}$ is defined as a mixture process whose conditional variance possesses a common GARCH component and a FI-GARCH component, i.e. the cumulative distribution function of e_t conditional on the past information has the form of

$$F(x|\mathcal{F}_{t-1}) = \alpha G(xh_{1,t}^{-1/2}) + (1-\alpha)G(xh_{2,t}^{-1/2}) \text{ for } x \in \mathbb{R}, \tag{2.1}$$

where $\mathcal{F}_t$ is the σ-field generated by $\{e_t, e_{t-1}, ...\}$, $0 < \alpha < 1$, $G(\cdot)$ is a cumulative distribution function with $\int xdG(x) = 0$ and $\int x^2 dG(x) = 1$,

$$h_{1,t} = \omega + \sum_{i=1}^{p} a_i e_{t-i}^2 + \sum_{j=1}^{q} b_j h_{1,t-j} \quad \text{and} \quad h_{2,t} = \frac{\gamma}{\beta(1)} + \left[1 - \frac{\delta(B)}{\beta(B)}(1-B)^d\right] e_t^2 \tag{2.2}$$

with $0 < d < 1$ and polynomials $\beta(x) = 1 - \sum_{i=1}^{m} \beta_i x^i$ and $\delta(x) = 1 - \sum_{i=1}^{s} \delta_i x^i$. Let $\{\varepsilon_t\}$ be independent and identically distributed (*i.i.d.*) random variables with cumulative distribution function $G(\cdot)$, and $\{z_t\}$ be *i.i.d.* Bernoulli random variables with $P(z_t = 1) = \alpha$. The MM-GARCH model (2.1) then can be represented as

$$e_t = \varepsilon_t \sqrt{h_t}, \quad h_t = z_t h_{1,t} + (1-z_t)h_{2,t}.$$

Note that $\{\varepsilon_t\}$ is the innovation sequence with mean zero and variance one, $\{z_t\}$ are latent variables, and they are independent of each other.

The conditional variances in (2.2) can be rewritten into the following ARCH(∞) forms,

$$h_{1,t} = c_0^{(1)} + \sum_{j=1}^{\infty} c_j^{(1)} e_{t-j}^2 \quad \text{and} \quad h_{2,t} = c_0^{(2)} + \sum_{j=1}^{\infty} c_j^{(2)} e_{t-j}^2, \tag{2.3}$$

i.e., the MM-GARCH model is a special mixture ARCH(∞) model, of which the two mixing components have different structures. One of them has geometric (short) memory volatility (as $c_j^{(1)}$ decay geometrically) while the other has hyperbolic (long) memory volatility (as $c_j^{(2)}$ decay hyperbolically).

Assumption 1. *$\omega > 0$, $a_i \geq 0$ with $i = 1, ..., p$, $b_j \geq 0$ with $j = 1, ..., q$ and $\sum_{i=1}^{p} a_i + \sum_{j=1}^{q} b_j < 1$.*

Assumption 2. *$c_0^{(2)} > 0$ and $c_j^{(2)} \geq 0$ for $j = 1, 2, \cdots$.*

Assumption 1 is the necessary and sufficient condition for the GARCH model to have a unique strictly stationary solution with finite variance. Assumption 2 is equivalent to the nonnegativity condition of the corresponding FIGARCH model, which has complicated forms even for models with lower orders. We can refer to Baillie et al. (1996), Bollerslev and Mikkeslen (1996), Chung (1999) and Conrad and Haag (2006) for details. Based on the above two assumptions, we can derive the following theorem, and the proof is relegated to the Appendix.

Theorem 1. *Under Assumptions 1 and 2, there exists a strictly stationary solution to* (2.1) *and* (2.2) *with finite variance, and such a solution is unique and non-anticipative.*

In order to study the higher order moments of e_t and h_t, we present another assumption as follows.

Assumption 3. *$c_0^{(1)} > 0$, $c_j^{(1)} \geq 0$ for $j = 1, 2, ...$ and $m_{2P}^{1/P} \cdot \sum_{j=1}^{\infty} [\alpha (c_j^{(1)})^P + (1 - \alpha)(c_j^{(2)})^P]^{1/P} < 1$, where $m_{2P} = \int x^{2P} dG(x) < \infty$ and P is a positive integer.*

Theorem 2. *Under Assumptions 2 and 3, there exists a unique, nonanticipative and strictly stationary solution to* (2.1) *and* (2.2) *with $E(h_t^P) < \infty$ and $E(e_t^{2P}) < \infty$.*

Theorem 2 provides a sufficient condition of the finite fourth moment for the mixture ARCH(∞) models with two variance components as in (2.3), which is more general than the MM-GARCH model. Note that, by Hölder's inequality, $m_{2P}^{1/P} \geq 1$ and $[\alpha (c_j^{(1)})^P + (1 - \alpha)(c_j^{(2)})^P]^{1/P} \geq \alpha c_j^{(1)} + (1 - \alpha) c_j^{(2)}$ for any positive integer j, i.e. the assumptions of Theorem 2 are stronger than those of Theorem 1.

Under Assumptions 1 and 2, it holds that $c_j^{(1)} = O(\rho^j)$ for some $\rho \in (0, 1)$, and $c_j^{(2)} = O(j^{-1-d})$. The coefficient $c_j^{(2)}$ will dominate $c_j^{(1)}$ when j is large, i.e. the long memory property of the MM-GARCH model is completely determined by the FIGARCH component. We sometimes may be interested in the autocovariance function of e_t^2. Under the assumptions of Theorem 2 with $P = 2$, it can be shown that $\mathrm{cov}(e_j^2, e_0^2) = O(j^{-1-d})$ by following the method in Giraitis and Surgailis (2002). It is noteworthy to point out that $\sum_{j=1}^{\infty} |\mathrm{cov}(e_j^2, e_0^2)| < \infty$.

The assumption that the mixing proportion α is a constant may exclude many real situations, and some exogenous variables may also affect the prediction and description of the target time series, see Wong and Li (2001a). In order to make the proposed MM-GARCH model more flexible, we may consider a time varying mixture proportion, which can be specified by a logistic link function, and this leads to a dynamic MM-GARCH. To further take into account the conditional mean structure, we introduce a general dynamic MM-GARCH process,

$$F(y|\mathcal{F}_{t-1}, \Omega_t) = \alpha_t G\left(\frac{y-\mu_t}{\sqrt{h_{1,t}}}\right) + (1-\alpha_t) G\left(\frac{y-\mu_t}{\sqrt{h_{2,t}}}\right) \quad \text{for } y \in \mathbb{R}, \tag{2.4}$$

where $\{y_t\}$ is the observed time series with structure $y_t = \mu_t + e_t$ and extra parameter vector θ_μ, $h_{1,t}$ and $h_{2,t}$ are defined as in (2.2), $\mathbf{x}_t = (x_{1t}, ..., x_{kt})'$ consists of k exogenous variables, Ω_t is the σ-field generated by $\{\mathbf{x}_t, \mathbf{x}_{t-1}, ...\}$ and

$$\log\left(\frac{\alpha_t}{1-\alpha_t}\right) = r_t = \lambda_0 + \sum_{i=1}^{l} \lambda_i (y_{t-i} - \mu_{t-i}) + \mathbf{x}_t' \varphi \tag{2.5}$$

with $\varphi = (\varphi_1, ..., \varphi_k)'$ being a k-dimensional parameter vector. Note that there also exist other specifications of the link function depending on practical concerns. Comparing to the dynamic MM-GARCH model, the one with constant mixing proportion is called the constant MM-GARCH model.

3 Estimation and the Observed Information Matrix

Denote the parameter vector of the dynamic MM-GARCH model (2.4) and (2.5) by $\theta = (\theta_\mu', \theta_{1h}', \theta_{2h}', \theta_\alpha')'$, where $\theta_{1h} = (\omega, a_1, ..., a_p, b_1, ..., b_q)'$, $\theta_{2h} = (\gamma, \beta_1, ..., \beta_m, \delta_1, ..., \delta_s, d)'$

and $\theta_\alpha = (\lambda_0, \lambda_1, ..., \lambda_l, \varphi_1, ..., \varphi_k)'$. Suppose that $y_1, ..., y_n$ are generated by the dynamic MM-GARCH model with exogenous variables $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$, and the true parameter vector θ_0 is an interior point of a compact set Θ. Define the functions $\mu_t(\theta_\mu)$, $h_{1,t}(\theta_\mu, \theta_{1h})$, $h_{2,t}(\theta_\mu, \theta_{2h})$ and $\alpha_t(\theta_\mu, \theta_\alpha)$ corresponding to equations (2.4) and (2.5), and for simplicity, we denote them respectively by $\mu_t(\theta)$, $h_{1,t}(\theta)$, $h_{2,t}(\theta)$ and $\alpha_t(\theta)$. In real applications, there are only n values available, however, the above functions all depend on infinite past observations. Hence, some initial values are needed, and we may simply assume $y_s = 0$ for $s \leq 0$. The effect of the initial values can be shown to be asymptotically negligible when $d > 0.5$, see Robinson and Zaffaroni (2006). Without loss of generality, we assume that the values y_s for $s \leq 0$ are all observable in the forthcoming discussions.

Assumption 4 (Identifiability condition)**.** *Polynomials* $1 - \sum_{j=1}^{q} b_j x^j$ *and* $\sum_{i=1}^{p} a_i x^i$ *have no common root, polynomials* $\beta(x) = 1 - \sum_{i=1}^{m} \beta_i x^i$ *and* $\delta(x) = 1 - \sum_{i=1}^{s} \delta_i x^i$ *have no common root, matrix* $E(\mathbf{x}_i \mathbf{x}_i')$ *has full rank, and* μ_t *satisfies some identifiability conditions.*

By temporarily assuming the normality on $G(\cdot)$ or ε_t, we can obtain the pseudo-log likelihood function

$$
\begin{aligned}
L^*(\theta) &= \sum_{t=1}^{n} \log \left\{ \frac{dF(y|\mathcal{F}_{t-1}, \Omega_t)}{dy}|_{y=y_t} \right\} \\
&= \sum_{t=1}^{n} \log \left\{ \frac{\alpha_t(\theta)}{\sqrt{h_{1,t}(\theta)}} \cdot \exp\left(\frac{[y_t - \mu_t(\theta)]^2}{2h_{1,t}(\theta)} \right) + \frac{1 - \alpha_t(\theta)}{\sqrt{h_{2,t}(\theta)}} \cdot \exp\left(\frac{[y_t - \mu_t(\theta)]^2}{2h_{2,t}(\theta)} \right) \right\} \\
&\quad - \frac{n}{2} \log(2\pi).
\end{aligned}
$$

Under Assumption 4, we employ the EM algorithm to find out the pseudo-maximum likelihood estimator, which is the maximizer of $L^*(\theta)$. By taking into account latent variables $\{z_t\}$, we treat $\{y_t, z_t, t = 1, ..., n\}$ as the complete data, and the complete pseudo-log likelihood function has the form of

$$
L(\theta) = \sum_{t=1}^{n} l_t(\theta),
$$

where

$$
\begin{aligned}
l_t(\theta) =& z_t \log[\alpha_t(\theta)] + (1 - z_t) \log[1 - \alpha_t(\theta)] \\
&- 0.5 z_t \left\{ \log[h_{1,t}(\theta)] + \frac{[y_t - \mu_t(\theta)]^2}{h_{1,t}(\theta)} \right\} - 0.5(1 - z_t) \left\{ \log[h_{2,t}(\theta)] + \frac{[y_t - \mu_t(\theta)]^2}{h_{2,t}(\theta)} \right\}.
\end{aligned}
$$

Correspondingly, the function $L^*(\theta)$ is referred to the incomplete likelihood function.

For simplicity, denote $z_{1,t} = z_t$, $z_{2,t} = 1 - z_t$, $\alpha_{1,t} = \alpha_t$ and $\alpha_{2,t} = 1 - \alpha_t$. Then the first order derivatives of $L(\theta)$ with respect to the parameters in each component are given as follows:

$$
\begin{aligned}
\frac{\partial L(\theta)}{\partial \theta_\mu} &= \sum_{t=1}^{n} \sum_{k=1}^{2} z_{k,t} \left[\frac{1}{\alpha_{k,t}} \frac{\partial \alpha_{k,t}}{\partial \theta_\mu} + \frac{1}{2h_{k,t}} \frac{\partial h_{k,t}}{\partial \theta_\mu} \left(\frac{e_t^2}{h_{k,t}} - 1 \right) - \frac{e_t}{h_{k,t}} \frac{\partial e_t}{\partial \theta_\mu} \right], \\
\frac{\partial L(\theta)}{\partial \theta_{kh}} &= \sum_{t=1}^{n} z_{k,t} \left[\frac{1}{2h_{k,t}} \frac{\partial h_{k,t}}{\partial \theta_{kh}} \left(\frac{e_t^2}{h_{k,t}} - 1 \right) \right], \qquad k = 1, 2 \\
\frac{\partial L(\theta)}{\partial \theta_\alpha} &= \sum_{t=1}^{n} (z_{1,t} - \alpha_{1,t}) \frac{\partial r_t}{\partial \theta_\alpha},
\end{aligned}
$$

where $e_t = y_t - \mu_t$. $\partial e_t / \partial \mu_t$ and $\partial r_t / \partial \theta_\alpha$ depend on the particular specifications of μ_t and r_t for different setups about the real data. The partial derivatives of $h_{k,t}$ with respect to the parameters in each conditional variance component are:

$$
\begin{aligned}
\frac{\partial h_{1,t}}{\partial \omega} &= 1 + \sum_{j=1}^{q} b_j \frac{\partial h_{1,t-j}}{\partial \omega}, \\
\frac{\partial h_{1,t}}{\partial a_i} &= e_{t-i}^2 + \sum_{j=1}^{q} b_j \frac{\partial h_{1,t-j}}{\partial a_i}, \quad i = 1, \cdots, p \\
\frac{\partial h_{1,t}}{\partial b_k} &= h_{1,t-k} + \sum_{j=1}^{q} b_j \frac{\partial h_{1,t-j}}{\partial b_k}, \quad j = 1, \cdots, q \\
\frac{\partial h_{2,t}}{\partial \gamma} &= 1 + \sum_{j=1}^{m} \beta_j \frac{\partial h_{2,t-j}}{\partial \gamma}, \\
\frac{\partial h_{2,t}}{\partial \delta_i} &= (1 - B)^d e_{t-i}^2 + \sum_{j=1}^{m} \beta_j \frac{\partial h_{2,t-j}}{\partial \delta_i}, \qquad i = 1, \cdots, s \\
\frac{\partial h_{2,t}}{\partial \beta_k} &= -e_{t-k}^2 + h_{2,t-k} + \sum_{j=1}^{m} \beta_j \frac{\partial h_{2,t-j}}{\partial \beta_k}, \qquad k = 1, \cdots, m
\end{aligned}
$$

$$\frac{\partial h_{2,t}}{\partial d} = -\delta(B)(1-B)^d log(1-B)e_t^2 + \sum_{j=1}^{m} \beta_j \frac{\partial h_{2,t-j}}{\partial d}.$$

The iterative EM procedure (Dempster et al., 1977) has been demonstrated readily flexible for estimating the parameters in the mixture-type models, as well as the mixture time series models. We describe these two steps as follows.

E-step: Suppose that the true parameter vector θ_0 is known. Then we can replace the missing data $\{z_{k,t}\}$ by their conditional expectations (denote by $\tau_{k,t}$), conditional on the parameters and the observed data $\{y_t\}$ and $\{\mathbf{x}_t\}$. Then, similar to Wong and Li (2001a, b), we can obtain the E-equation as follows,

$$\tau_{k,t} = \frac{\alpha_{k,t} h_{k,t}^{-1/2} g([y_t - \mu_t] h_{k,t}^{-1/2})}{\sum_{j=1}^{2} \alpha_{j,t} h_{k',t}^{-1/2} g([y_t - \mu_t] h_{j,t}^{-1/2})}, \quad \text{for} \quad k = 1, 2, \tag{3.1}$$

where $g(\cdot)$ is the density function of the standard normal distribution. Obviously, $\tau_{1,t} = 1 - \tau_{2,t}$.

M-step: Suppose the missing data $\{z_{k,t}\}$ are known by replacing them with $\{\tau_{k,t}\}$. The estimates of the parameter vector θ can then be obtained by maximizing the likelihood function $L(\theta) = \sum_{t=1}^{n} l_t(\theta)$, hence

$$\widehat{\theta}_n = \underset{\theta \in \Theta}{\operatorname{argmax}}\, L(\theta).$$

We can then obtain the estimates of the parameter vector θ by iterating these two steps until convergence. When there is no conditional mean part, i.e. $\mu_t = 0$, the optimization in the M-step can be simplified to the following separate maximizations:

$$\widehat{\theta}_{1h} = \operatorname{argmin} \sum_{t=1}^{n} z_t \left\{ \log[h_{1,t}(\theta)] + \frac{y_t^2}{h_{1,t}(\theta)} \right\}, \tag{3.2}$$

$$\widehat{\theta}_{2h} = \operatorname{argmin} \sum_{t=1}^{n} (1 - z_t) \left\{ \log[h_{2,t}(\theta)] + \frac{y_t^2}{h_{2,t}(\theta)} \right\}, \tag{3.3}$$

and

$$\widehat{\theta}_\alpha = \text{argmax} \sum_{t=1}^{n} \{z_t \log[\alpha_t(\theta)] + (1 - z_t) \log[1 - \alpha_t(\theta)]\}. \tag{3.4}$$

By the missing information principle in Hartley and Hocking (1971), the observed information matrix I can be calculated by

$$I = I_c - I_m = E\left(-\frac{\partial^2 L(\theta)}{\partial\theta\partial\theta'}|\theta, \mathcal{F}_n, \Omega_n\right)_{\widehat{\theta}_{EM}} - \text{Var}\left(\frac{\partial L(\theta)}{\partial\theta}|\theta, \mathcal{F}_n, \Omega_n\right)_{\widehat{\theta}_{EM}}, \tag{3.5}$$

where $\widehat{\theta}_{EM}$ is the parameter estimate from the EM procedure and it should be close to the true parameter vector θ_0, see also Woodbury (1971); I_c is the complete information matrix and I_m is the missing information matrix. It is known that if the distribution of the innovations is symmetric, the information matrix is approximately block diagonal. Hence the standard errors of the parameter estimators in each component can be calculated separately. This method has been used in the papers aforementioned. The details of equation (3.5) are put in the Appendix.

Note that, since latent variables $\{z_{k,t}\}$ are regime indicators and are conditionally independent, we may calculate the observed information matrix alternatively by

$$I = \sum_{t=1}^{n} J_t(\widehat{\theta}_{EM}) J_t'(\widehat{\theta}_{EM}), \tag{3.6}$$

where

$$\begin{aligned} J_t(\theta) &= E\left(\frac{\partial l_t(\theta)}{\partial\theta}|\theta, \mathcal{F}_n, \Omega_n\right) \\ &= \sum_{k=1}^{2} \tau_{k,t} \left\{\frac{1}{\alpha_{k,t}}\frac{\partial\alpha_{k,t}}{\partial\theta} + \frac{y_t - \mu_t}{h_{k,t}}\frac{\partial\mu_t}{\partial\theta} + \frac{1}{2h_{k,t}}\left[\frac{(y_t - \mu_t)^2}{h_{k,t}} - 1\right]\frac{\partial h_{k,t}}{\partial\theta}\right\}. \end{aligned} \tag{3.7}$$

The justification of equations (3.6) and (3.7) can be found in Louis (1982). Compared with equation (3.5), equation (3.6) only involves the first derivative and guarantees I to be always non-negative definite.

The asymptotic standard errors of parameter estimators $\widehat{\theta}_{EM}$ are the square roots of the diagonal elements of the inverse of their observed information matrix defined above. However, when ε_t is not normally distributed, the asymptotic variance of $\widehat{\theta}_{EM}$ may depend on other quantities besides the information matrix, and then its standard errors are not available. We leave it for possible future research.

4 Model Selection Criterion

In this section, we discuss the problem of model selection for the MM-GARCH models proposed in Section 2, and the likelihood function is usually involved. Note that the complete log-likelihood $L(\theta)$ in Section 3 contains the unobservable random variables $\{z_t\}$. Alternatively, we use the incomplete likelihood $L^*(\theta)$, and selection criteria based on $L^*(\theta)$ may have better performance than those based on $L(\theta)$, see Wong and Li (2000).

Based on the above likelihood function, the Bayesian information criterion (BIC) for the dynamic MM-GARCH models can be defined as

$$\text{BIC} = -2L^*(\widehat{\theta}_{EM}) + \log(n) \cdot (D_\mu + p + q + m + s + l + k + 4),$$

where D_μ is the dimension of θ_μ, and the fitted model with the minimum BIC value is selected. We conduct a small simulation experiment to evaluate its performance, and the data generating process with the dynamic mixture proportion (Model 2) in the next section are employed. Note that the true values of the orders for the common GARCH component, the FIGARCH component and the mixture proportion component are $(p, q) = (1, 1)$, $(m, s) = (1, 0)$ and $(l, k) = (1, 0)$, respectively. The orders (m, s) in the FIGARCH component are fixed since, when they increase, the constraints will be very restrictive such that the EM procedure becomes intractable. We consider three different values for each of the other orders: $p, q, l = 1$, 2 or 3 and $k = 0$, 1 or 2. The exogenous variables $\{\mathbf{x}_t\}$ are independent AR(1) processes, i.e. $x_{1,t} = 0.6x_{1,t-1} + \varepsilon_{1,t}$, $x_{2,t} = -0.2x_{2,t-1} + \varepsilon_{2,t}$, and $(\varepsilon_{1,t}, \varepsilon_{2,t})'$ are *i.i.d.* random vectors with the standard multivariate normal distribution. There are then 81 candidate models. We generate 200 replications with sample size 2500, and the estimating procedure in the next section is used. For each replication, the values of BIC are calculated for all candidate models, and the model with the smallest BIC is selected. As a result, the BIC correctly identified the true orders in all components at a rate of 98%. What's more, all 200 replications have correctly identified p, q and k, only 4 replication misidentified l.

Remember that we proposed two models in Section 2, the constant and dynamic MM-GARCH. It is natural to ask whether the dynamic mixture proportion is necessary for a real data set, and it can be formalized into the following hypothesis

$$H_0 : \lambda_1 = \cdots = \lambda_l = \varphi_1 = \cdots = \varphi_k = 0.$$

The likelihood ratio test statistic for testing the null hypothesis H_0 against its simple negation is $-2(L_0^* - L_1^*)$, where L_0^* and L_1^* are the maximized log-likelihood $L^*(\theta)$ respectively under the null and the alternative hypotheses, and it can be shown to have an asymptotic standard χ^2 distributions with $l + k$ degrees of freedom. Another simulation experiment is conducted to evaluate its empirical sizes as in Wong and Li (2001a). We employ the same data generating process as that in the next section with the mixture proportion specified by

$$\alpha_{1,t} = \frac{1}{1 + \exp(0.7 - \lambda_1 y_{t-1} - \varphi_1 x_t)},$$

where $x_t = 0.6x_{t-1} + \varepsilon_t$, $\varepsilon_t \sim N(0, 1)$. Two likelihood ratio tests are considered as follows,

(i) $H_0 : \lambda_1 = 0$ assuming $\varphi_1 = 0$; and

(ii) $H_0 : \lambda_1 = \varphi_1 = 0$.

The sample size is set to $n = 2500$ or 5000, and there are 200 replications for each sample size. We consider three commonly used significance levels, 1%, 5% and 10%, and the empirical sizes are listed in Table 1. It can be seen that empirical sizes match their nominal rates very well even when the sample size is 2500.

5 Simulation Experiments

Monte Carlo simulation experiments are conducted to investigate the finite sample performance of the EM algorithm in MM-GARCH models and the theoretical standard errors based on Louis (1982).

Since we concentrate on the behavior of the second moment, the mean structure μ_t in all the experiments is assumed to be zero. Thus $y_t = e_t$. Then, we consider the logistic link function which is only dependent on y_t and no exogenous variables are included for simplicity. Now the parameter vector $\theta = (\theta'_{1h}, \theta'_{2h}, \theta'_{\alpha})'$.

The data are generated from the following process:

$$\begin{cases} y_t = \varepsilon_t \sqrt{h_t}, \quad h_t = z_{1,t} h_{1,t} + (1 - z_{1,t}) h_{2,t}, \\ h_{1,t} = 0.1 + 0.3 y_{t-1}^2 + 0.25 h_{1,t-1}, \\ h_{2,t} = 0.45 + [(1 - 0.2B) - (1 - B)^{0.8}] y_t^2 + 0.2 h_{2,t-1}, \end{cases}$$

where $\varepsilon_t \sim i.i.d.N(0,1)$ and $P(z_{1,t} = 1) = \alpha_{1,t}$. One component of the volatility specification is a GARCH$(1,1)$ with the parameter set $\theta_{1h} = (\omega, a_1, b_1) = (0.1, 0.3, 0.25)$ and the other is a FIGARCH$(1, d, 0)$ with the parameter set $\theta_{2h} = (\gamma, d, \beta_1) = (0.45, 0.8, 0.2)$. Note that the parameter combination in each volatility component should guarantee the nonnegativity of the corresponding conditional variance. Here we impose positive constrains on ω, a_1, b_1 of the GARCH parameters and $\gamma > 0$, $0 < \beta_1 \leq d < 1$ of the FIGARCH parameters. The latter condition is from Chung (1999). The mixture proportion $\alpha_{1,t}$ is given respectively by:

Model 1: constant mixture proportion

$$\alpha_{1,t} = \frac{1}{1 + \exp(-\lambda_0)}, \quad \text{with} \quad \lambda_0 = -0.7;$$

or

Model 2: dynamic mixture proportion, specified by a logistic link function

$$\alpha_{1,t} = \frac{1}{1 + \exp(-\lambda_0 - \lambda_1 y_{t-1})}, \quad \text{with} \quad (\lambda_0, \lambda_1) = (-0.7, 0.4).$$

For each model, the sample sizes are $n = 2500$, 5000, or 10000 and there are 1000 independent replications. For each replication, we generate $n+5000$ data points, and then discard the first 5000 ones to mitigate the effect of the initial values. In the FIGARCH component, we truncate 200 terms of $\{\pi_j\}$ in the expansion of $(1-B)^d = 1 - \sum_{j=1}^{\infty} \pi_j B^j$ without any significant influence to the final results.

The estimation method we employ here is the traditional EM algorithm and it is somewhat time consuming. One may refer to other improved EM-type algorithms to speed up the convergence rate. Recall the EM algorithm in Section 3, we know that in the E-step, the estimation of the unobservable $z_{k,t}$ in the i-th iteration is updated by

$$\tau_{1,t}^{(i)} = \frac{\hat{\alpha}_{1,t}^{(i-1)} \phi(y_t/\sqrt{\hat{h}_{1,t}^{(i-1)}})/\sqrt{\hat{h}_{1,t}^{(i-1)}}}{\hat{\alpha}_{1,t}^{(i-1)} \phi(y_t/\sqrt{\hat{h}_{1,t}^{(i-1)}})/\sqrt{\hat{h}_{1,t}^{(i-1)}} + \hat{\alpha}_{2,t}^{(i-1)} \phi(y_t/\sqrt{\hat{h}_{2,t}^{(i-1)}})/\sqrt{\hat{h}_{2,t}^{(i-1)}}},$$

where $\phi(\cdot)$ is the density function of the standard normal distribution. $\hat{h}_{k,t}^{(i-1)} = h_{k,t}(\widehat{\theta}_{EM}^{(i-1)})$ and $\widehat{\theta}_{EM}^{(i-1)}$ is the estimation of θ in the $(i-1)th$ iteration. Then substituting $z_{1,t}$ by $\tau_{1,t}^{(i)}$ in equations (3.2)-(3.4), we obtain the maxima $\widehat{\theta}_{1h}^{(i)}$, $\widehat{\theta}_{2h}^{(i)}$ and $\widehat{\theta}_{\alpha}^{(i)}$, and hence $\widehat{\theta}_{EM}^{(i)}$. The iteration stops when $|\widehat{\theta}_{EM}^{(i)} - \widehat{\theta}_{EM}^{(i-1)}| \leq 10^{-6}$.

Tables 2 and 3 give the biases, empirical standard errors (MSE) and two theoretical standard errors (ASE1 and ASE2), respectively under Models 1 and 2. The values of ASE1 are calculated from (3.5), and those of ASE2 are from (3.6). As it is known that

$$\frac{\partial^2 l_t}{\partial \theta_{1h} \partial \theta_{2h}} = \frac{\partial^2 l_t}{\partial \theta_{kh} \partial \theta_{\alpha}} = 0, \quad \text{for} \quad k = 1, 2,$$

in (3.5), the parameters are divided into three parts and the observed information matrix $I \simeq \mathrm{diag}(I^{(\theta_{1h})}, I^{(\theta_{2h})}, I^{(\theta_{\alpha})})$. The standard errors of parameter estimators in each part are calculated as the square roots of the diagonal elements of the inverse of individual observed information matrices.

In general, for both models, the estimators have small biases and MSE even when n is 2500, which is a relatively small under the situation where both mixture regimes and hyperbolic decaying ARCH(∞) coefficients are involved. Remarkably, the values of ASE2 in both tables are closer to those of MSE than those of ASE1. It is under expectation since: first, separately calculating the observed information matrix for the parameters in each part may result in some biases as the whole observed information matrix is not exactly block diagonal; second, the second order derivatives are involved in (3.5) and the bias of the approximation in equation (A.7) may not be small when n is not large enough. Hence we use equation (3.6) to calculate the theoretical standard errors. It is also noted that the difference between ASE2 and MSE becomes small when the sample size is 10000 at both models. Moreover, the empirical and theoretical standard errors have approximately $\sqrt{n}$-convergence rates, which match the theoretical asymptotic.

6 Real Data Analysis

In this application, we consider the daily S&P500 index data from January 2, 1990 to April 27, 2012 with 5628 observations. The log return series is defined as $y_t = \log(P_t) - \log(P_{t-1})$, where P_t is the closing price at day t. We target at the centralized y_t. The original price series and y_t (DLprice) are plotted in Figure 1. The series clearly shows the occurrence of tranquil and volatile periods. The sample period covers the Asian Financial Crisis in 1997, the outburst of IT Bubble in 2001, the Severe Acute Respiratory Syndrome (SARS) in 2003 and the Global Financial Crisis (GFC) in 2007. We divide the whole sample into two parts. The first part is from January 2, 1990 to May 5, 2010 with 5128 observations and we use this subsample as in-sample data to conduct model estimation. The second part is from May 6, 2010 to April 27, 2012 with 500 observations and we use it as out-of-sample data to perform model forecasting.

We employ three different volatility models to fit the structure of the second order dependence. These volatility models are GARCH(1,1), HYGARCH(1,d_v,1) and dynamic MM-GARCH. We use the dynamic MM-GARCH model directly rather than the constant MM-GARCH model since it is more general. The two mixture components in the dynamic MM-GARCH model are respectively the GARCH(1,1) and the FIGARCH(1, d_v, 0). The estimation results are given in Table 4. The theoretical standard errors of parameter estimates are reported in the parentheses.

In the fitted GARCH(1, 1) model, the value of $\hat{a}_1 + \hat{b}_1$ is very close to 1, which is the canonical IGARCH effect. In the second model, the estimator of the variance indicator $\hat{\alpha}$ in the HYGARCH(1, d_v, 1) being sightly greater than 1 but not significantly different from 1, also reveals the possible non-covariance-stationarity of the process. In the last model, the first component of the volatility is found to be a stable GARCH ($\hat{a}_1 + \hat{b}_1 = 0.95 < 1$) while the second one is a FIGARCH and always non-stationary. In both HYGARCH and dynamic MM-GARCH models, the estimates of parameters satisfy the nonnegativity conditions for the conditional variance, see Conrad and Haag (2006) and Conrad (2010). Remarkably, $\hat{\lambda}_1$ in the dynamic MM-GARCH seems a little bigger (15.1266), this is because we use the return series itself instead of the series in percentage. Among these three models, the dynamic MM-GARCH model has both the largest value of the log likelihood function and the smallest value of BIC, suggesting that it has the best fitting performance for this data set. The likelihood ratio test statistic for testing $H_0 : \lambda_1 = 0$ equals 7.4, significantly exceeding the 5% critical value of χ^2_1 (3.84), indicating also that the dynamic mixture proportion is necessary.

Figure 2 plots the two components in the conditional variance of the fitted dynamic MM-GARCH model. It is shown that the FIGARCH component mimics more dramatic variations than the GARCH during the whole period, especially during the periods of different crisis. Figure 3 represents the dynamic mixture proportion of variance falling into the short memory GARCH component. Note that the link function adopted here is $\log(\alpha_{1,t}/1 - \alpha_{1,t}) = \lambda_0 + \lambda_1 |y_{t-1}|$. It is reported in the graph that $\alpha_{1,t}$ is around 0.25 and the largest fluctuation occurs during the period of the GFC.

Finally, we consider the out-of-sample coverage rate of lower and upper 95% predictive intervals under the three volatility models. With the idea of Value-at-Risk, we also calculate the corresponding unconditional coverage test of Kupiec (1995) and the conditional coverage test of Christoffersen (1998). The predictive intervals are obtained with rolling estimation. The results of one-step-ahead and five-step-ahead forecasts are respectively reported in Tables 5 and 6. It is seen that all the three models pass both of the tests. Compared to the GARCH model, the HYGARCH and dynamic MM-GARCH have higher p-values on the lower tails, which implies that the failure rate of exceeding the worst return are more likely to be correctly specified under these two models. Although there are no significant evidence to show that HYGARCH and dynamic MM-GARCH have better performance than GARCH when the forecast horizon is increased beyond 5 days, as suggested by Conrad (2010), we may still expect both of them to have more accurate predictive performances when the forecast horizons are long enough since the long memory volatility models have been shown to possess definite superiority in long term forecasts.

7 Conclusion

In this paper, we propose the mixture memory GARCH models to describe the dynamics of the volatility of financial returns. Between the two mixture components, one is the short memory common GARCH and the other is the long memory FIGARCH. The mixture probability lying in [0,1] makes it possible for a mixture memory GARCH model to share both properties of covariance stationarity and long memory volatility. Using the Volterra series expansion, we obtain the sufficient conditions for the existence of stationary solution to the constant MM-GARCH model with P-th order moment (P is a positive integer). We also discuss the EM algorithm as a parameter estimation procedure and provide a simple formula for calculating the observed information matrix. Asymptotic standard errors of the estimators are thus calculated more easily and accurately than those in some previous papers, see Wong and Li (2001a,b) and Cheng et al. (2009). A model selection criterion and a likelihood based mis-specification test are also briefly discussed. Finally, the real application to S&P 500 data set illustrates the competitiveness of our models.

Acknowledgement

We thank the editor, an associate editor and three anonymous referees for their valuable comments that led to a substantial improvement of the paper, and the Hong Kong Research Grants Council grant HKU 702908P for partial support.

Appendix: Technical Details

This appendix gives the proofs of Theorems 1 and 2 and the detailed form of equation (3.5) in calculating the observed information matrix.

Proof of Theorem 1. It is sufficient to show the result for the following difference equation,

$$e_t^2 = \varepsilon_t^2 h_t, \quad h_t = z_t h_{1,t} + (1 - z_t) h_{2,t}, \tag{A.1}$$

where $h_{1,t}$ and $h_{2,t}$ are defined as that in (2.2). This can be proved by using a Volterra series type expansion (Priestley, 1988) of the above equation, and the method is similar to that in Giraitis et al. (2000).

Denote $h_{1,t} = c_0^{(1)} + \sum_{j=1}^{\infty} c_j^{(1)} e_{t-j}^2$ and $h_{2,t} = c_0^{(2)} + \sum_{j=1}^{\infty} c_j^{(2)} e_{t-j}^2$. Let $\psi_j(t) = z_t c_j^{(1)} +$

$(1-z_t)c_j^{(2)}$ for $j = 0, 1, \ldots$. By the iterative equation (A.1), we have that

$$
\begin{aligned}
e_t^2 &= \psi_0(t)\varepsilon_t^2 + \sum_{j_1=1}^{\infty} \psi_{j_1}(t)\varepsilon_t^2 e_{t-j_1}^2 \\
&= \psi_0(t)\varepsilon_t^2 + \sum_{j_1=1}^{\infty} \psi_{j_1}(t)\psi_0(t-j_1)\varepsilon_t^2\varepsilon_{t-j_1}^2 + \sum_{j_1=1}^{\infty}\sum_{j_2=1}^{\infty} \psi_{j_1}(t)\psi_{j_2}(t-j_1)\varepsilon_t^2\varepsilon_{t-j_1}^2 e_{t-j_1-j_2}^2 \\
&= \cdots \\
&= \psi_0(t)\varepsilon_t^2 + \sum_{l=1}^{\infty} \sum_{j_1,\ldots,j_l=1}^{\infty} \psi_{j_1}(t)\psi_{j_2}(t-j_1)\cdots\psi_{j_l}(t-j_1-\cdots-j_{l-1}) \\
&\qquad\qquad \psi_0(t-j_1-\cdots-j_l)\varepsilon_t^2\varepsilon_{t-j_1}^2\cdots\varepsilon_{t-j_1-\cdots-j_l}^2. \qquad \text{(A.2)}
\end{aligned}
$$

It is easy to verify that the above process is a solution to (A.1). Denote $\psi_j = E[\psi_j(t)] = \alpha c_j^{(1)} + (1-\alpha)c_j^{(2)}$. It holds that, by Assumption 1, $\sum_{j=1}^{\infty} c_j^{(1)} < 1$. Since $\sum_{j=1}^{\infty} c_j^{(2)}$ is always 1, then $\rho = \sum_{j=1}^{\infty} \psi_j < 1$. Note that $E(\varepsilon_t^2) = 1$. Hence,

$$
E(e_t^2) = \psi_0 + \sum_{l=1}^{\infty} \sum_{j_1,\ldots,j_l=1}^{\infty} \psi_{j_1}\psi_{j_2}\cdots\psi_{j_l}\psi_0 = \frac{\psi_0}{1-\rho} < \infty,
$$

i.e., the stochastic process $\{e_t^2\}$ in (A.2) is well defined.

We next show the uniqueness, and assume that $\{x_t^2\}$ is another strictly stationary solution to (A.1) with $E(x_t^2) < \infty$. Then by the iterative equation (A.2), after M (M is a positive number) steps, we have that

$$
\begin{aligned}
x_t^2 =& \psi_0(t)\varepsilon_t^2 + \sum_{l=1}^{M} \sum_{j_1,\ldots,j_l=1}^{\infty} \psi_{j_1}(t)\psi_{j_2}(t-j_1)\cdots\psi_{j_l}(t-j_1-\cdots-j_{l-1})\psi_0(t-j_1-\cdots-j_l) \\
&\qquad \cdot \varepsilon_t^2\varepsilon_{t-j_1}^2\cdots\varepsilon_{t-j_1-\cdots-j_l}^2 \\
&+ \sum_{j_1,\ldots,j_{M+1}=1}^{\infty} \psi_{j_1}(t)\psi_{j_2}(t-j_1)\cdots\psi_{j_{M+1}}(t-j_1-\cdots-j_M) \\
&\qquad \cdot \varepsilon_t^2\varepsilon_{t-j_1}^2\cdots\varepsilon_{t-j_1-\cdots-j_M}^2 x_{t-j_1-\cdots-j_{M+1}}^2,
\end{aligned}
$$

and

$$
\begin{aligned}
x_t^2 - e_t^2 = &\sum_{j_1,\ldots,j_{M+1}=1}^{\infty} \psi_{j_1}(t)\psi_{j_2}(t-j_1)\cdots\psi_{j_{M+1}}(t-j_1-\cdots-j_M) \\
&\qquad\qquad\qquad \cdot \varepsilon_t^2\varepsilon_{t-j_1}^2\cdots\varepsilon_{t-j_1-\cdots-j_M}^2 x_{t-j_1-\cdots-j_{M+1}}^2 \\
&- \sum_{l=M+1}^{\infty}\sum_{j_1,\ldots,j_l=1}^{\infty} \psi_{j_1}(t)\psi_{j_2}(t-j_1)\cdots\psi_{j_l}(t-j_1-\cdots-j_{l-1})\psi_0(t-j_1-\cdots-j_l) \\
&\qquad\qquad\qquad \cdot \varepsilon_t^2\varepsilon_{t-j_1}^2\cdots\varepsilon_{t-j_1-\cdots-j_l}^2. \qquad (A.3)
\end{aligned}
$$

For the first term in the above (A.3), by Chebyshev's inequality, and the assumption of unanticipativity as well as $Ex_t^2 < \infty$,

$$
\begin{aligned}
&\forall \varepsilon > 0, \\
&\varepsilon \cdot P\{\sum_{j_1,\ldots,j_{M+1}=1}^{\infty} \psi_{j_1}(t)\psi_{j_2}(t-j_1)\cdots\psi_{j_{M+1}}(t-j_1-\cdots-j_M) \\
&\qquad\qquad\qquad \cdot \varepsilon_t^2\varepsilon_{t-j_1}^2\cdots\varepsilon_{t-j_1-\cdots-j_M}^2 x_{t-j_1-\cdots-j_{M+1}}^2 > \varepsilon\} \\
&\leq E|\sum_{j_1,\ldots,j_{M+1}=1}^{\infty} \psi_{j_1}(t)\psi_{j_2}(t-j_1)\cdots\psi_{j_{M+1}}(t-j_1-\cdots-j_M) \\
&\qquad\qquad\qquad \cdot \varepsilon_t^2\varepsilon_{t-j_1}^2\cdots\varepsilon_{t-j_1-\cdots-j_M}^2 x_{t-j_1-\cdots-j_{M+1}}^2| \\
&= E(x_t^2)\cdot \sum_{j_1,\cdots,j_{M+1}} \psi_{j_1}\cdots\psi_{j_{M+1}} \\
&= E(x_t^2)\cdot \rho^{M+1}.
\end{aligned}
$$

Then by the Borel-Cantelli lemma, it converges to zero almost surely.

For the second term at (A.3), we denote

$$
\begin{aligned}
\zeta_t(l, j_1, \ldots, j_l) =&\varepsilon_t^2\varepsilon_{t-j_1}^2\cdots\varepsilon_{t-j_1-\cdots-j_l}^2 \\
&\cdot \psi_{j_1}(t)\psi_{j_2}(t-j_1)\cdots\psi_{j_l}(t-j_1-\cdots-j_{l-1})\psi_0(t-j_1-\cdots-j_l).
\end{aligned}
$$

Then $\sum_{l=0}^{\infty}\zeta_t(l, j_1, \cdots, j_l) < \infty$ a.s. if $\rho < 1$. Let $M \to \infty$, the second term can be made arbitrarily small with probability 1. Thus, $x_t^2 = e_t^2$ in the almost surely sense. $\square$

Proof of Theorem 2. It is sufficient to show that $E(e_t^{2P}) < \infty$, $P \geq 2$ since Assumptions 2 and 3 imply the assumptions in Theorem 1. We first show the case with $P = 2$.

Recall that

$$\begin{aligned}\zeta_t(l, j_1, ..., j_l) =& \varepsilon_t^2 \varepsilon_{t-j_1}^2 \cdots \varepsilon_{t-j_1-\cdots-j_l}^2 \\ & \cdot \psi_{j_1}(t)\psi_{j_2}(t-j_1)\cdots\psi_{j_l}(t-j_1-\cdots-j_{l-1})\psi_0(t-j_1-\cdots-j_l).\end{aligned}$$

Then, by the proof of Theorem 1,

$$e_t^2 = \psi_0(t)\varepsilon_t^2 + \sum_{l=1}^{\infty}\sum_{j_1,...,j_l=1}^{\infty}\zeta_t(l, j_1, ..., j_l).$$

Denote

$$\psi_j^{(2)} = E[\psi_j^2(t)] = \alpha(c_j^{(1)})^2 + (1-\alpha)(c_j^{(2)})^2 \quad \text{for} \quad j \geq 0.$$

Remember that $E(\varepsilon_t^4) = m_4$. It holds that

$$E[\zeta_t(l, j_1, ..., j_l)]^2 = m_4^{l+1}\psi_{j_1}^{(2)}\cdots\psi_{j_l}^{(2)}\psi_0^{(2)}.$$

Hence, by Hölder's inequality,

$$\begin{aligned}& E\left[\sum_{l=1}^{\infty}\sum_{j_1,...,j_l=1}^{\infty}\zeta_t(l, j_1, ..., j_l)\right]^2 \\ & = \sum_{l=1}^{\infty}\sum_{j_1,...,j_l=1}^{\infty}\sum_{l^*=1}^{\infty}\sum_{j_1^*,...,j_l^*=1}^{\infty} E[\zeta_t(l, j_1, ..., j_l)\zeta_t(l^*, j_1^*, ..., j_l^*)] \\ & \leq \sum_{l=1}^{\infty}\sum_{j_1,...,j_l=1}^{\infty}\sum_{l^*=1}^{\infty}\sum_{j_1^*,...,j_l^*=1}^{\infty} m_4^{(l+1)/2}(\psi_{j_1}^{(2)}\cdots\psi_{j_l}^{(2)}\psi_0^{(2)})^{1/2} m_4^{(l^*+1)/2}(\psi_{j_1^*}^{(2)}\cdots\psi_{j_l^*}^{(2)}\psi_0^{(2)})^{1/2} \\ & = m_4\psi_0^{(2)}\left(\frac{\rho^*}{1-\rho^*}\right)^2 < \infty,\end{aligned}$$

where $\rho^* = m_4^{1/2}\sum_{j=1}^{\infty}(\psi_j^{(2)})^{1/2} < 1$. We finish the proof for the case $P = 2$.

For a general positive integer P, by Hölder's inequality again,

$$E|\prod_{i=1}^{P} X_i| \leq \prod_{i=1}^{P}\left(E|X_i|^P\right)^{1/P}, \quad \text{for} \quad P \geq 1$$

where $X_i,\ i=1,\cdots,P$ are real- or complex-valued random variables. Then we have that

$$
\begin{aligned}
&E\left[\sum_{l=1}^{\infty}\sum_{j_1,\cdots,j_l=1}^{\infty}\zeta_t(l,j_1,\cdots,j_l)\right]^P\\
=&\sum_{l_1=1}^{\infty}\sum_{j_1,\cdots,j_{l_1}=1}^{\infty}\cdots\sum_{l_P=1}^{\infty}\sum_{j_1,\cdots,j_{l_P}=1}^{\infty}E\left[\zeta_t(l_1,j_1,\cdots,j_{l_1})\cdots\zeta_t(l_P,j_1,\cdots,j_{l_P})\right]\\
\leq&\sum_{l_1=1}^{\infty}\sum_{j_1,\cdots,j_{l_1}=1}^{\infty}\cdots\sum_{l_P=1}^{\infty}\sum_{j_1,\cdots,j_{l_P}=1}^{\infty}\left\{\prod_{i=1}^{P}E[\zeta_t(l_i,j_1,\cdots,j_{l_i})]^P\right\}^{1/P}. \qquad (A.4)
\end{aligned}
$$

Denote $\psi_j^{(P)}=E[\psi_j(t)]^P=\alpha[c_j^{(1)}]^P+(1-\alpha)[c_j^{(2)}]^P$ for $j\geq 0$. It holds that

$$
E[\zeta_t(l_i,j_1,\cdots,j_{l_i})]^P=m_{2P}^{l_i+1}\psi_{j_1}^{(P)}\cdots\psi_{j_{l_i}}^{(P)}\psi_0^{(P)},
$$

and

$$
\left\{\prod_{i=1}^{P}E[\zeta_t(l_i,j_1,\cdots,j_{l_i})]^P\right\}^{1/P}=m_{2_P}\psi_0^{(P)}\prod_{i=1}^{P}m_{2P}^{l_i/P}[\psi_{j_1}^{(P)}\cdot\psi_{j_{l_i}}^{(P)}]^{1/P}, \qquad (A.5)
$$

where $m_{2P}=E(\varepsilon_t^{2P})$. Finally, based on (A.5) and the last term in (A.4),

$$
\begin{aligned}
&E\left[\sum_{l=1}^{\infty}\sum_{j_1,\cdots,j_l=1}^{\infty}\zeta_t(l,j_1,\cdots,j_l)\right]^P\\
&\leq m_{2_P}\psi_0^{(P)}\left\{\sum_{l_1=1}^{\infty}\sum_{j_1,\cdots,j_{l_1}=1}^{\infty}\left[m_{2P}^{1/P}[\psi_{j_1}^{(P)}]^{1/P}\cdots m_{2P}^{1/P}[\psi_{j_{l_1}}^{(P)}]^{1/P}\right]\right\}^P\\
&=m_{2_P}\psi_0^{(P)}\left[\frac{\eta}{1-\eta}\right]^P, \qquad (A.6)
\end{aligned}
$$

where $\eta=m_{2P}^{1/P}\sum_{j=1}^{\infty}[\psi_j^{(P)}]^{1/P}<1$. We finish the proof of Theorem 2. $\square$

Detailed form of equation (3.5). We calculate the complete and the missing information matrices separately. Denote by $I_c^{(\theta_{1h})}$, $I_c^{(\theta_{2h})}$ and $I_c^{(\theta_\alpha)}$ respectively the complete information matrix *w.r.t.* θ_{1h},θ_{2h} and θ_α. Then, denote by $I_m^{(\theta_{1h})}$, $I_m^{(\theta_{2h})}$ and $I_m^{(\theta_\alpha)}$ respectively the missing information matrix *w.r.t.* θ_{1h},θ_{2h} and θ_α.

Note that

$$\frac{\partial^2 L(\theta)}{\partial\theta_{kh}\partial\theta'_{kh}} = \sum_{t=1}^{n} z_{k,t}\left[\frac{1}{2h_{k,t}^2}\frac{\partial h_{k,t}}{\partial\theta_{kh}}\frac{\partial h_{k,t}}{\partial\theta'_{kh}}\left(\frac{2e_t^2}{h_{k,t}}-1\right)-\frac{1}{2h_{k,t}}\frac{\partial^2 h_{k,t}}{\partial\theta_{kh}\partial\theta'_{kh}}\left(\frac{e_t^2}{h_{k,t}}-1\right)\right],$$
$$\text{for} \quad k=1,2$$

As $E(z_{t,k}\big|\theta,\mathcal{F}_n,\Omega_n)=\tau_{k,t}$, for any $1\le t\le n$, the complete information matrix for θ_{kh} are approximately as follows

$$I_c^{(\theta_{kh})} = -E\left(\frac{\partial^2 L(\theta)}{\partial\theta_{kh}\partial\theta'_{kh}}\big|\theta,\mathcal{F}_n,\Omega_n\right)_{\widehat{\theta}_{EM}} \simeq \sum_{t=1}^{n}\tau_{k,t}\left[\frac{1}{2h_{k,t}^2}\frac{\partial h_{k,t}}{\partial\theta_{kh}}\frac{\partial h_{k,t}}{\partial\theta'_{kh}}\right]_{\widehat{\theta}_{EM}}, \quad k=1,2. \tag{A.7}$$

For the missing information matrix of θ_{kh},

$$\begin{aligned} I_m^{(\theta_{kh})} &= \mathrm{Var}\left(\frac{\partial L(\theta)}{\partial\theta}\big|\theta,\mathcal{F}_n,\Omega_n\right)_{\widehat{\theta}_{EM}} \\ &= E\left(\frac{\partial L_c(\theta)}{\partial\theta_{kh}}\frac{\partial L_c(\theta)}{\partial\theta'_{kh}}\big|\theta,\mathcal{F}_n,\Omega_n\right)_{\widehat{\theta}_{EM}} \\ &- \left[E\left(\frac{\partial L(\theta)}{\partial\theta_{kh}}\big|\theta,\mathcal{F}_n,\Omega_n\right)\right]_{\widehat{\theta}_{EM}}\left[E\left(\frac{\partial L(\theta)}{\partial\theta'_{kh}}\big|\theta,\mathcal{F}_n,\Omega_n\right)\right]_{\widehat{\theta}_{EM}}. \end{aligned} \tag{A.8}$$

Since

$$\begin{aligned} \frac{\partial L_c(\theta)}{\partial\theta_{kh}}\frac{\partial L_c(\theta)}{\partial\theta'_{kh}} &= \left[\sum_{t=1}^{n} z_{k,t}\frac{1}{2h_{k,t}}\frac{\partial h_{kt}}{\partial\theta_{kh}}\left(\frac{e_t^2}{h_{k,t}}-1\right)\right]\left[\sum_{t=1}^{n} z_{k,t}\frac{1}{2h_{k,t}}\frac{\partial h_{kt}}{\partial\theta_{kh}}\left(\frac{e_t^2}{h_{k,t}}-1\right)\right] \\ &= \sum_{t=1}^{n} z_{k,t}^2\frac{1}{4h_{k,t}^2}\frac{\partial h_{k,t}}{\partial\theta_{kh}}\frac{\partial h_{k,t}}{\partial\theta'_{kh}}\left(\frac{e_t^2}{h_{k,t}}-1\right)^2 \\ &+ 2\sum_{1\le t_i<t_j\le n} z_{k,t_i}z_{k,t_j}\frac{1}{4h_{k,t_i}h_{k,t_j}}\frac{\partial h_{k,t_i}}{\partial\theta_{kh}}\frac{\partial h_{k,t_j}}{\partial\theta'_{kh}}\left(\frac{e_{t_i}^2}{h_{k,t_i}}-1\right)\left(\frac{e_{t_j}^2}{h_{k,t_j}}-1\right). \end{aligned}$$

Substituting the equation above to (A.8), again by the conditional independence of z_{k,t_i} and z_{k,t_j} for $i\neq j$, we have

$$I_m^{(\theta_{kh})} = \sum_t[\tau_{k,t}-\tau_{k,t}^2]\frac{1}{4h_{k,t}^2}\frac{\partial h_{k,t}}{\partial\theta_{kh}}\frac{\partial h_{k,t}}{\partial\theta'_{kh}}\left(\frac{e_t^2}{h_{k,t}}-1\right)^2_{\widehat{\theta}_{EM}}.$$

For the parameters θ_α in the logistic link function, the complete and missing information matrices are respectively given as:

$$I_c^{(\theta_\alpha)} = \sum_{t=1}^{n} \left[\alpha_{k,t}(1-\alpha_{k,t}) \frac{\partial r_t}{\partial \theta_\alpha} \frac{\partial r_t}{\partial \theta'_\alpha} \right]_{\widehat{\theta}_{EM}},$$

$$I_m^{(\theta_\alpha)} = \sum_{t=1}^{n} \left[\tau_{1,t}(1-\tau_{1,t}) \frac{\partial r_t}{\partial \theta_\alpha} \frac{\partial r_t}{\partial \theta'_\alpha} \right]_{\widehat{\theta}_{EM}}.$$

Then the observed information matrices *w.r.t.* θ_{kh} and θ_α are calculated separately by

$$I^{(\theta_{kh})} = I_c^{(\theta_{kh})} - I_m^{(\theta_{kh})} \quad \text{and} \quad I^{(\theta_\alpha)} = I_c^{(\theta_\alpha)} - I_m^{(\theta_\alpha)}.$$

□

References

Baillie, R. T., T. Bollerslev, and H. O. Mikkeslen (1996). Fractionally integrated generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics 74*, 3–30.

Baillie, R. T. and C. Morana (2009). Modelling long memory and structural breaks in conditional variances: an adaptive figarch approach. *Journal of Economic Dynamics and Control 33*(8), 1577–1592.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics 31*, 307–327.

Bollerslev, T. and H. O. Mikkeslen (1996). Modeling and pricing long memory in stock market volatility. *Journal of Econometrics 73*, 151–184.

Cheng, X. X., L. H. Yu, and W. K. Li (2009). On a dynamic mixture garch models. *Journal of Forecasting 28*, 247–265.

Christoffersen, P. (1998). Evaluating interval forecasts. *International economic review 39*, 841–862.

Chung, C. (1999). Estimating the fractionally integrated garch model. *Working paper, National Taiwan University*.

Conrad, C. (2010). Non-negativity conditions for the hyperbolic garch model. *Journal of Econometrics 157*(2), 441–457.

Conrad, C. and B. Haag (2006). Inequality constraints in the fractionally integrated garch model. *Journal of Financial Econometrics 4*(3), 413–449.

Davidson, J. (2004). Moment and memory properties of linear conditional heteroscedasticity models, and a new model. *Journal of Business & Economic Statistics 22*, 16–29.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B 39*(1), 1–38.

Ding, Z. X., C. W. J. Grange, and R. F. Engle (1993). A long memory property of stock market returns and a new model. *Journal of Empirical Finance 1*, 83–106.

Ding, Z. X. and C. W. J. Granger (1996). Modeling volatility persistence of speculative returns: a new approach. *Journal of Econometrics 73*, 185–215.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica 50*, 987–1007.

Giraitis, L., P. Kokoszka, and R. Leipus (2000). Stationary arch models: dependence structure and central limit theorem. *Econometric Theory 16*, 3–22.

Giraitis, L. and D. Surgailis (2002). ARCH-type bilinear models with double long memory. *Stochastic Processes and their Applications 100*(1), 275–300.

Hartley, H. and R. Hocking (1971). The analysis of incomplete data. *Biometrics 72*, 783–823.

Kupiec, P. (1995). Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives 3*(2), 73–84.

Li, M., G. Li, and W. K. Li (2011). Score tests for hyperbolic garch models. *Journal of Business & Economic Statistics 29*(4), 579–586.

Louis, T. A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society, Series B 44*(2), 226–233.

Mikosch, T. and C. Starica (2004). Nonstationarities in financial time series, the long-range dependence, and the IGARCH effects. *Review of Economics and Statistics 86*(1), 378–390.

Priestley, M. B. (1988). *Non-linear and non-stationary time series analysis.* London: Academic Press.

Robinson, P. M. (1991). Testing for strong serial correlation and dynamic conditional heteroscedasticity in multiple regression. *Journal of Econometrics 47*(1), 67–84.

Robinson, P. M. and P. Zaffaroni (2006). Pseudo-maximum likelihood estimation of arch(∞) models. *The Annals of Statistics 34*, 1049–1074.

Wong, C. S. and W. K. Li (2000). On a mixture autoregressive model. *Journal of the Royal Statistical Society, Series B 62*(1), 95–115.

Wong, C. S. and W. K. Li (2001a). On a logistic mixture autoregressive model. *Biometrika 88*(3), 833–846.

Wong, C. S. and W. K. Li (2001b). On a mixture autoregressive conditional heteroscedastic model. *Journal of the American Statistical Association 96*(455), 982–995.

Woodbury, M. (1971). Discussion of paper by hartley and hocking. *Biometrics 27*, 808–817.

Zaffaroni, P. (2004). Stationarity and memory of arch(∞) models. *Econometric Theory 20*, 147–160.

Table 1: Empirical sizes of two likelihood ratio tests.

	$n = 2500$			$n = 5000$		
Test	1%	5%	10%	1%	5%	10%
(i)	0.015	0.060	0.095	0.010	0.055	0.105
(ii)	0.025	0.055	0.100	0.015	0.050	0.100

Table 2: Estimation results for the constant MM-GARCH model (Model 1).

	True	n=2500				n=5000				n=10000			
	Value	Bias	MSE	ASE1	ASE2	Bias	MSE	ASE1	ASE2	Bias	MSE	ASE1	ASE2
						GARCH component							
ω	0.10	0.0034	0.0492	0.0359	0.0410	0.0021	0.0343	0.0258	0.0296	0.0005	0.0218	0.0180	0.0207
a_1	0.30	0.0052	0.0989	0.0709	0.0886	0.0035	0.0703	0.0515	0.0645	-0.0002	0.0465	0.0362	0.0454
b_1	0.25	0.0157	0.1252	0.0991	0.1042	0.0074	0.0858	0.0723	0.0760	0.0039	0.0572	0.0510	0.0539
						FIGARCH component							
γ	0.45	0.0067	0.0877	0.0629	0.0851	0.0033	0.0582	0.0417	0.0552	-0.0027	0.0345	0.0279	0.0365
β_1	0.20	0.0113	0.1016	0.0966	0.1055	0.0059	0.0724	0.0657	0.0694	0.0021	0.0471	0.0437	0.0458
d	0.80	0.0086	0.0965	0.0933	0.1006	0.0037	0.0663	0.0623	0.0648	0.0032	0.0439	0.0408	0.0421
						Mixture proportion							
λ_0	-0.70	0.0767	0.4171	0.1332	0.3689	0.0367	0.2843	0.0936	0.2587	0.0094	0.1661	0.0647	0.1791

Table 3: Estimation results for the dynamic MM-GARCH model (Model 2).

	True	n=2500				n=5000				n=10000			
	Value	Bias	MSE	ASE1	ASE2	Bias	MSE	ASE1	ASE2	Bias	MSE	ASE1	ASE2
						GARCH component							
ω	0.10	-0.0019	0.0485	0.0351	0.0412	-0.0009	0.0314	0.0257	0.0299	-0.0027	0.0205	0.0180	0.0211
a_1	0.30	-0.0058	0.0784	0.0590	0.0748	-0.0049	0.0540	0.0431	0.0542	-0.0074	0.0363	0.0307	0.0389
b_1	0.25	0.0165	0.1160	0.0973	0.1039	-0.0089	0.0861	0.0715	0.0751	-0.0040	0.0544	0.0511	0.0548
						FIGARCH component							
γ	0.45	-0.0048	0.0838	0.0584	0.0740	-0.0055	0.0538	0.0398	0.0500	-0.0095	0.0356	0.0275	0.0344
β_1	0.20	0.0051	0.1062	0.0935	0.0998	0.0011	0.0742	0.0649	0.0673	0.0015	0.0483	0.0458	0.0471
d	0.80	0.0105	0.0958	0.0920	0.0974	0.0042	0.0692	0.0629	0.0644	0.0031	0.0440	0.0440	0.0446
						Mixing proportion							
λ_0	-0.70	-0.0711	0.4285	0.1473	0.3590	-0.0677	0.2658	0.1008	0.2506	-0.0763	0.1869	0.0706	0.1764
λ_1	0.40	0.0615	0.2124	0.1592	0.1877	0.0285	0.1208	0.1077	0.1252	0.0147	0.0869	0.0749	0.0864

Table 4: S&P500 from 01/03/1994-05/05/2010, sample size =5128

Parameters	GARCH	HYGARCH	DMM-GARCH
$\omega^{(1)}$	7.5558e-07		0.0000
	(1.0509e-007)		(0.0000)
a_1	0.0653		0.0181
	(0.0045)		(0.0050)
b_1	0.9291		0.9319
	(0.0048)		(0.0149)
$\omega^{(2)}$		4.4895e-006	4.2342e-006
		(1.2572e-006)	(6.3146e-007)
δ_1		0.1658	
		(0.0461)	
β_1		0.5825	0.8295
		(0.0642)	(0.0214)
α		1.0158	
		(0.0306)	
d_v		0.4439	0.8805
		(0.0620)	(0.0309)
λ_0			-1.1307
			(0.1595)
λ_1			15.1266
			(7.7842)
Log-likelihood Value	$1.6665e+004$	$1.6671e+004$	$1.6699e+004$
BIC	$-3.3304e+004$	$-3.3299e+004$	$-3.3329e+004$

Table 5: Comparison of forecasting performance based on the one-day ahead 95% predictive interval coverage rate.

	Out-of-sample unconditional coverage test statistic and p-value		Out-of-sample conditional coverage test statistic and p-value		Out-of-sample coverage probability(%)	
Model	Lower tail	Upper tail	Lower tail	Upper tail	Lower tail	Upper tail
GARCH(1,1)	2.4933	0.0385	2.5259	2.4648	93.2	95.2
	(0.1143)	(0.8444)	(0.2828)	(0.2916)		
HYGARCH$(1, d, 1)$	0.6592	0.3817	1.0751	2.4118	94	95.6
	(0.4169)	(0.5367)	(0.5842)	(0.2994)		
DMM-GARCH	1.4386	0.1729	2.1080	3.2636	93.6	94.6
	(0.2304)	(0.6776)	(0.3485)	(0.1956)		

Table 6: Comparison of forecasting performance based on the five-day ahead 95% predictive interval coverage rate.

	Out-of-sample unconditional coverage test statistic and p-value		Out-of-sample conditional coverage test statistic and p-value		Out-of-sample coverage probability(%)	
Model	Lower tail	Upper tail	Lower tail	Upper tail	Lower tail	Upper tail
GARCH(1,1)	1.5435	0.0242	2.1618	2.4711	93.75	95.16
	(0.2141)	(0.8765)	(0.3393)	(0.2907)		
HYGARCH$(1, d, 1)$	0.2094	0.6291	0.3985	2.4905	94.56	95.77
	(0.6472)	(0.4277)	(0.8193)	(0.2879)		
DMM-GARCH	0.4317	0.1333	0.5462	2.3757	94.35	95.36
	(0.5111)	(0.7151)	(0.7610)	(0.3049)		

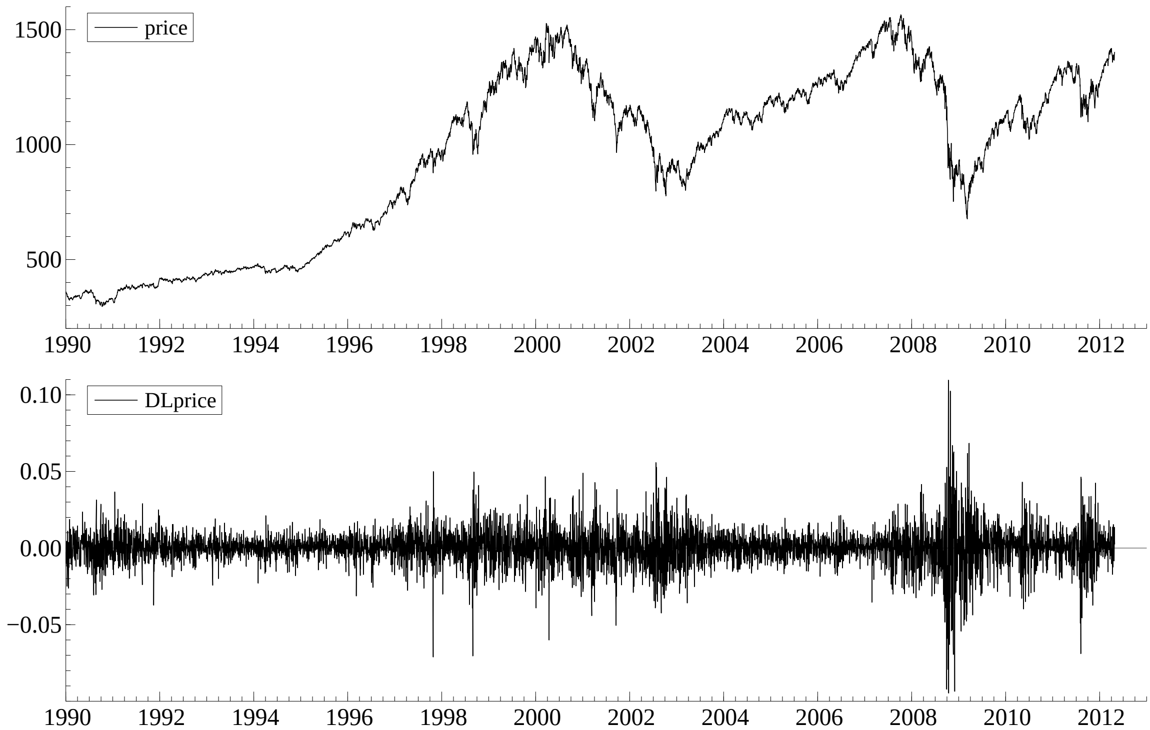

Figure 1: Daily closing prices (Price) of S&P500 Index and its log returns (DLprice) from January 2, 1990 to April 27, 2012.

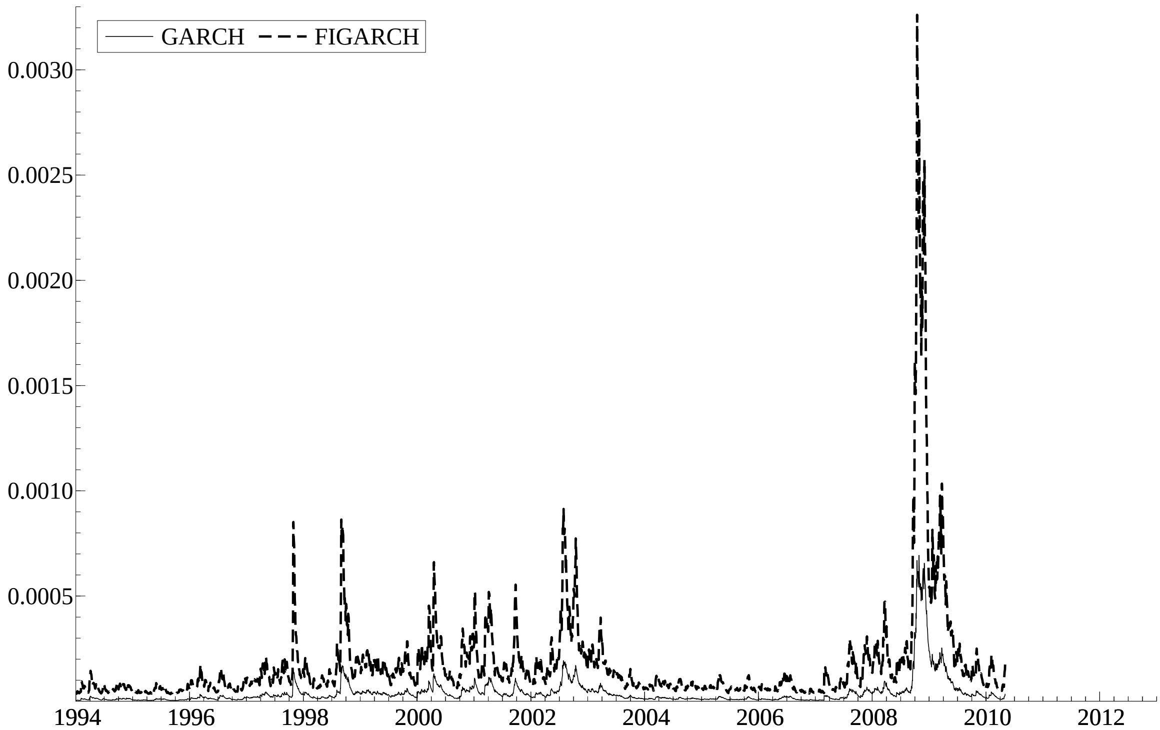

Figure 2: Fitted volatility components. The higher line represents the FIGARCH component while the lower one represents the GARCH component.

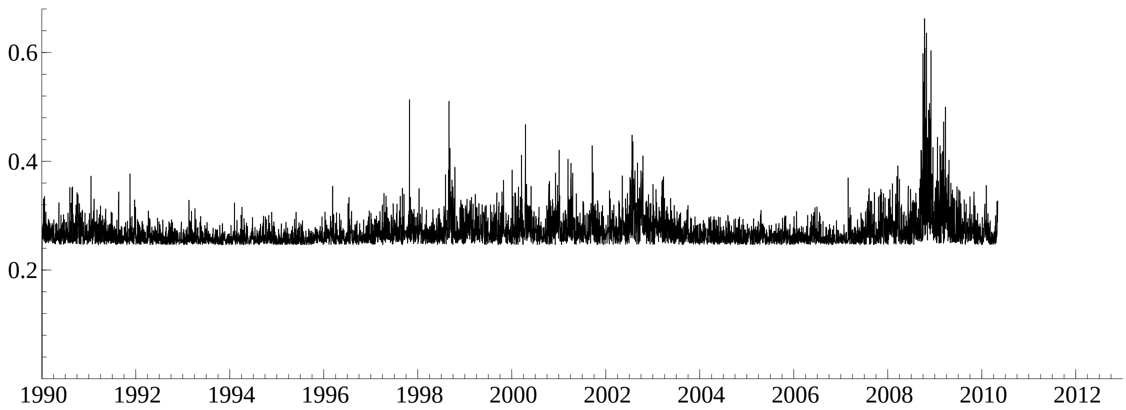

Figure 3: The dynamic mixture proportion which is the probability of volatility falling into the GARCH component.